MODERN SPEECH SYNTHESIS FOR PHONETIC SCIENCES: A DISCUSSION AND AN EVALUATION

Zofia Malisz[1], Gustav Eje Henter[1], Cassia Valentini-Botinhao[2], Oliver Watts[2], Jonas Beskow[1], Joakim Gustafson[1]

[1]KTH Royal Institute of Technology, Stockholm, Sweden, [2]The University of Edinburgh, UK
{malisz, ghe, beskow, jkgu}@kth.se, {cvbotinh, owatts}@inf.ed.ac.uk

ABSTRACT

Decades of gradual advances in speech synthesis have recently culminated in exponential improvements fuelled by deep learning. This quantum leap has the potential to finally deliver realistic, controllable, and robust synthetic stimuli for speech experiments. In this article, we discuss these and other implications for phonetic sciences. We substantiate our argument by evaluating classic rule-based formant synthesis against state-of-the-art synthesisers on a) subjective naturalness ratings and b) a behavioural measure (reaction times in a lexical decision task). We also differentiate between text-to-speech and speech-to-speech methods. Naturalness ratings indicate that all modern systems are substantially closer to natural speech than formant synthesis. Reaction times for several modern systems do not differ substantially from natural speech, meaning that the processing gap observed in older systems, and reproduced with our formant synthesiser, is no longer evident. Importantly, some speech-to-speech methods are nearly indistinguishable from natural speech on both measures.

Keywords: Speech synthesis, scientific methodology, speech technology

1. INTRODUCTION

The artificial modelling of human speech depends on an ongoing dialogue between phoneticians and engineers. Indeed, speech science helped synthesis get started [18]: phonetics was instrumental in speech processing and engineering in the formant synthesis age, when data was sparse and modelling took place in wetware rather than software. Likewise, in today's data-driven speech technology with algorithms and machine learning, perception-based modelling, such as the mel scale, is standard. Also, advanced evaluation methods crossed over from perceptual phonetics to text-to-speech (TTS) and benchmark TTS robustness and precision.

As in any dialogue, the influence is reciprocal. Milestones in phonetic sciences, such as evidencing categorical speech perception, were reached with the use of synthetic sound continua [21]. Consequently, theoretical advances such as the motor theory of speech perception [20] or acoustic cue analysis were made possible by experiments with synthetic stimuli. Natural speech contains redundant and residual cues to place of articulation which are difficult to exclude in, e.g., manipulation of formant transitions using natural speech [3]. Synthetic stimuli offered control over single-cue variability limiting confounds, making it viable to assess listeners' sensitivity to a particular acoustic cue in isolation.

Arguably, control over numerous meaningful, relatively low-level signal properties such as pitch, VOT, etc. has been the central feature of rule-based formant synthesis (henceforth: *classical speech synthesis*) in phonetic research. Conversely, the inability of concatenative signal generation methods to create a continuum of acoustic cues in response to input control has excluded these from much such research. The one notable exception is MBROLA [8], which uses a waveform-modification technique similar to PSOLA [25] to allow control of pitch and duration given a sequence of allophones to speak. Applications include speech distortion and delexicalisation, empirical paradigms where cues to particular structures, such as prosody, are removed.

TTS continues to provide tools and heuristics for the everyday phonetic business (stimuli creation) but it also offers whole modelling frameworks used for testing phonological models (analysis by synthesis) [41, 5].

Unfortunately, the many differences in human perception between natural and classical synthesised speech have cast doubt on the universality of research findings reliant on synthetic speech [15]. Classical synthesis has proven to be generally less intelligible and to overburden attention and cognitive mechanisms resulting in slower processing times [7]. Winters & Pisoni [39] present a comprehensive review of such effects and explain them by showing that stimuli from formant synthesis are "impoverished" in terms of perceptual cues.

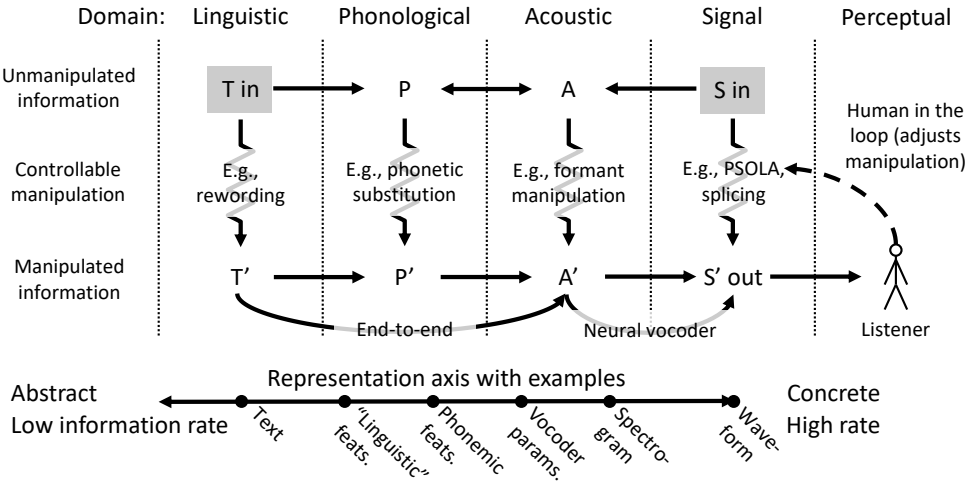
2. WHAT CAN MODERN SYNTHESIS DO?

Having pointed out some of the traditional uses, benefits, and drawbacks of using classical synthesis in phonetic research, we now discuss how *modern speech synthesis* (statistical parametric and end-to-end TTS approaches plus neural vocoders) can address these points in terms of output control and realism. We also consider the resources needed for state-of-the-art performance. The discussion is closely tied to the overview of speech technology for phonetic research presented in Figure 1.

2.1. Control

The vast majority of synthesis systems have an input that allows control over the phoneme sequence expressed by the system. The input can be either text or phones (like in text-to-speech) or speech (like in voice conversion).

Figure 1: A unifying view of speech representations (capital letters), optional transformations (horizontal arrows), and controllable manipulations (vertical arrows). The main inputs are shown on shaded backgrounds. The representation axis along the bottom shows the domains and abstract-to-concrete ordering of common speech representations.



These key inputs are shown shaded in Figure 1. The input is then transformed between different intermediate representations by processing steps (arrows in the figure) to eventually produce a speech output waveform. However, a phonetician often requires the ability to manipulate additional aspects of the output at various levels of abstraction. Such controllable manipulations are illustrated by squiggly vertical arrows in Figure 1.

Recent speech technology solutions concentrate on learning rather than designing methods for speech output control. Hence, the manipulation arrows in Figure 1 are realised through mappings learned by supervised machine learning. In principle, this enables control of arbitrary concepts that are hard to define acoustically, such as speaker identity, age, and gender [23], emotional state [12], and prosodic prominence [24]. These approaches can easily be adapted to learn to control other characteristics such as formant frequencies.

There are no guarantees that a particular type of control is learnable, although the listed successes suggest great potential for speech perception applications. Additionally, learned control can be made more precise by including a human in the loop who adjusts the control inputs until the output is satisfactory, as seen in Figure 1.

2.2. Realism

Improving the realism of the generated output is of prime concern for applications in speech research to avoid the issues listed in [39]. However, it is difficult to exhaustively define the necessary cues to make synthesis realistic, e.g., natural, easily intelligible, and specific to a given speaker. Substantial advances in realism have therefore come from improved signal processing and, particularly, from including ever more machine learning and statistical modelling into the speech processing steps (the various arrows in Figure 1). Intelligibility in quiet, a prominent issue with formant synthesis [39], has been on par with natural speech since the era of hidden Markov models and decision trees in TTS [17]. Replacing decision trees with neural networks subsequently improved the rated speech quality of these systems [38].

Since 2016, important breakthroughs have come from introducing deep learning for generating audio waveforms, i.e., neural vocoders like WaveNet [26], and for analysing text input (Tacotron [36]). These innovations have radically increased the mean opinion score (MOS) quality ratings of text-to-speech [31] and speech-to-speech [22]. The best results tend to come from simultaneously learning to perform as many processing steps as possible, known as *end-to-end* synthesis (cf. Figure 1).

Speech waveforms have a much higher information rate than text, since they also express speaker and channel properties, emphasis, and so on. To generate speech signals from text requires filling in missing information as one progresses along the representation axis in Figure 1. This is not easy, and realism is oftentimes higher when stimuli are generated from high-rate speech audio input (like voice conversion) rather than low-rate text input (like TTS). Perhaps for this reason, speech-in-speech-out (SISO) pipelines tend to be more prevalent than text-in-speech-out (TISO) pipelines in phonetic research. Hybrid approaches combining speech and text input (both of the shaded boxes in Figure 1) can be used to compensate for TTS weaknesses, e.g., by imposing prosodic characteristics extracted from natural speech onto TTS [11, 33].

2.3. Resource needs

Modern synthesis methods leverage learning to improve from increased resources, such as speech recordings, annotation, and computation power. Massive computational resources were a pre-requisite for breakthroughs such as WaveNet [26] and Tacotron 2 [31]. This performance scaling contrasts with classical speech synthesis and manipulation methods based on explicit modelling, which do not improve with additional resources.

However, the trade-off between results and resources

can be circumvented. Modern systems can mimic new speaker voices from limited adaptation data [23, 16], removing the need to record full TTS corpora for each new speaker. Extending these capabilities to under-resourced languages and language varieties seems possible, given more research into multilingual synthesis technology.

Speech technology trends that push towards making data and code fully available, and the dropping cost of computation, inspire confidence that resources will not be a limiting factor in the future. As done in image and text processing, speech technologists could share pre-trained controllable synthesis systems to facilitate wider adoption of these methods for phonetic research.

3. COMPARATIVE PARADIGM EVALUATIONS

We now describe an experiment to assess the decreased perceptual difference between natural speech and modern synthesis, relative to the differences described by Winters & Pisoni [39] for classical synthesis. We, therefore, compare classic synthesis (as a baseline) and modern synthesisers in terms of listener ratings and listener behaviour. This evaluation focuses on building TISO and SISO systems with realism levels representative of what can be achieved by open code and databases with modest computation, leaving pronunciation/output control as future work along with resource-demanding neural vocoders. Our test stimuli, anonymised responses, and analysis scripts can all be found at doi.org/10.7488/ds/2520.

3.1. Data for training and evaluation

We used a male RP British English speaker recordings [6], available at doi.org/10.7488/ds/2482, downsampled to 16 kHz using `sox -r 16k`. This data contains ca. 2000 sentences for system building, plus 720 Harvard and 300 modified rhyme test (MRT) utterances, detailed in Section 3.4, that we set aside for evaluations.

3.2. Systems

We introduce the systems we evaluate below and in Table 1. All test stimuli were normalised to the same active speech level using ITU P.56.

NAT Natural speech recordings held out from training.

VOC Copy synthesis (acoustic analysis followed by re-synthesis) with the MagPhase vocoder [9]. MagPhase is a high-fidelity signal generator that, for better quality, retains both magnitude and phase spectrum information.

MERLIN Synthetic speech generated by the Merlin TTS system [40] using the MagPhase vocoder. Specifically, this system uses feed-forward deep neural networks trained to predict F0, magnitude, and phase from so-called linguistic features extracted from text using the Festival [35] front-end with the Combilex [28, 29] RP British English dictionary. This system represents state-of-the-art research-grade statistical parametric TTS.

GL Copy synthesis from magnitude mel-spectrograms using the Griffin-Lim algorithm [10] for phase reconstruction. This is a classic technique from signal processing that has seen a resurgence with end-to-end systems as a simple and fast baseline signal generator.

Table 1: System labels, input and output feature types, modelling paradigms, and signal generation methods for the systems compared in this study. Inputs and outputs are SI: speech in, TI: text in, SO: speech out.

System	Type	Paradigm	Signal gen.
NAT	-	Natural	Vocal tract
VOC	SISO	Copy synthesis	MagPhase
MERLIN	TISO	Stat. parametric	MagPhase
GL	SISO	Copy synthesis	Griffin-Lim
DCTTS	TISO	End-to-end	Griffin-Lim
OVE	TISO	Rule-based	Formant

DCTTS Text-to-speech using deep convolutional networks as in [34] with Griffin-Lim signal generation. Like Tacotron [36], DCTTS is closer than Merlin to full end-to-end speech synthesis. Tacotron-like systems combined with neural vocoders can produce highly realistic speech output [31], surpassing all other TTS paradigms. Similar to [33, 37], instead of graphemes, we used Combilex phonetisations of words as input, retaining punctuation to enable synthesising more appropriate intonation. For accurate pronunciation, DCTTS was trained both on our speaker and on about 11,600 utterances from another English speaker from [19]. The last network in the model was fine-tuned to adapt to the target voice.

OVE The multilingual rule-based formant TTS system [4] (OVE III), implemented as a Tcl module by [32] and configured to use a male RP British English voice. This system is representative of research-grade formant-based TTS. Unlike other systems, OVE permits optional prosodic emphasis control of individual output words.

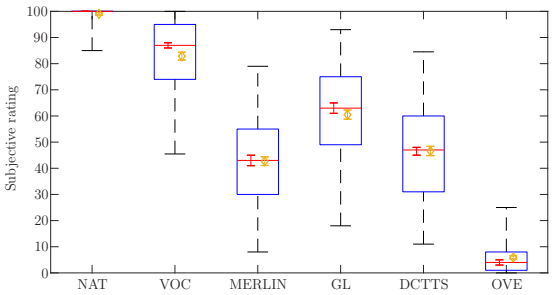
3.3. Subjective rating experiment

To assess realism we use a setup closely resembling the ITU standard MUSHRA [14] (MUltiple Stimuli with Hidden Reference and Anchor). This test presents listeners with randomly ordered, unlabelled, parallel stimuli from different systems all speaking the same text. Listeners rate the *naturalness* of each stimulus against a designated natural reference (here NAT). They can listen to the stimuli in any order as many times as they like and adjust their ratings until they are satisfied. These tests have been found to resolve system differences better than traditional MOS evaluations [27]. The MUSHRA rating scale from 0 to 100 was anchored by assigning the standard MOS labels "Bad", "Poor", "Fair", "Good", and "Excellent" to successive 20-point intervals along the rating scale.

The test used 20 native English-speaking listeners (students at the University of Edinburgh) with no known hearing impairments. Listeners rated stimuli representing the different systems speaking four sets of ten Harvard sentences [30] (designed to be approximately phonetically balanced) and were remunerated for their efforts.

The box plot in Figure 2 visualises the 799 ratings for each system analysed from the listening test. All pairwise system differences are statistically significant at the 0.01 level ($p < 10^{-6}$) after Holm-Bonferroni correction for multiple comparisons, using either paired Wilcoxon signed-rank tests (for the median) or paired Student's t-

Figure 2: Box plot of per-system MUSHRA ratings. 95% confidence intervals are shown for medians (red box centre lines) and means (yellow diamonds). Whiskers extend to cover 95% of all responses.



tests (for the mean). Ignoring ties, VOC was rated above NAT 5.7% of the time. The same number for MERLIN was 0.38%, the highest for a TISO system. OVE was rated below any other system 99% of the time or more.

3.4. Lexical decision task

To assess intelligibility and processing speed via correct-response rate and a behavioural measure, namely, reaction time, we set up a lexical decision task using ExperimentMFC, an interface for forced-choice listening tests in Praat. Stimuli were CVC words from 50 minimal pairs selected from the modified rhyme test [13], embedded in a fixed carrier sentence rendered by the six different systems. We used emphasis control in OVE to approximate the NAT productions of the MRT word token.

We tested 20 listeners that were selected and remunerated as in the previous test. For each stimulus, listeners were asked to push one out of two buttons to indicate which word they heard out of the minimal pair, as soon as they knew. Listeners heard each system speak each of the 100 words once, producing 600 responses and reaction times per listener. Stimuli were ordered randomly for each listener while the order of the two response options was balanced across listeners in two blocks.

Table 2 reports the results of mixed-effects linear regression on logarithmic response-times with data trimming and model criticism based on [2]. Apart from OVE and DCTTS, reaction times to all synthetic systems do not exhibit statistically significant differences relative to NAT. This means that one state-of-the-art TISO and both SISO systems analysed here are processed similarly to natural speech by human listeners. Incorrect response rates (final column of Table 2) are comparable to NAT for VOC and MERLIN, with DCTTS and OVE showing the worst performance on this indicator of intelligibility.

4. DISCUSSION

We have experimentally verified that modern synthetic speech is perceptually closer to natural speech than are classical synthesis methods, to such an extent that modern methods largely overcome the processing inadequacies that have previously plagued synthesised stimuli in speech sciences. As the next step, we want to extend this

Table 2: Results of mixed-effects linear regression modelling on log reaction times and the percentage of incorrect responses in the lexical decision task.

System	Est. (±std. err.)	p-value	Incorrect
NAT (ref.)			2.6%
GL	-0.001 (±0.02)	= 0.94	4.0%
VOC	0.02 (±0.02)	= 0.33	2.5%
DCTTS	0.04 (±0.02)	< *0.01*	5.8%
MERLIN	0.02 (±0.02)	= 0.14	3.0%
OVE	0.09 (±0.02)	< *0.001*	6.0%

investigation to include speech manipulation and neural vocoders, and also consider more evaluation paradigms. We expect that this will confirm our hypothesis that modern speech synthesis can improve on the utility of classic phonetic research tools in many scenarios. While it is always an option to use formant synthesis if a quality of "artificial" speech is desired for an experiment, it seems that findings from synthetic stimuli that are as close as possible to natural speech should generalise better to natural speech perception for most tasks.

An interesting short-term technological goal is to create neural vocoders that are controlled by formants, phonological features, or other parameters frequently manipulated in phonetic research. This should permit speech manipulations in both a SISO framework (like [5]) and a TISO framework (like KlaTTStat [1]) while simultaneously improving realism. With enough resources, this paradigm might theoretically even surpass the realism attained by PSOLA when manipulating pitch and duration. The TISO approach would also benefit from end-to-end learning to better predict the chosen vocoder parameters, improving all aspects of the TISO pipeline.

More fanciful applications of synthesis technologies can also be envisioned. Good neural vocoders trained without any control input at all can generate highly realistic single-talker babble [26]. This seems like a promising tool for work on de-lexicalised speech. Another application is the study of optional or paralinguistic phenomena that stem from unplanned processes, such as realistic hesitations, backchannels, or non-phonemic conversational clicks. As natural examples of such phenomena are difficult to elicit from human speakers in empirical designs, the ability to synthesise these phenomena on demand would greatly benefit systematic study.

In the long run, we aim to demonstrate that new tools from speech technology enable novel scientific discovery in speech sciences, and ultimately advance research in the field. To do this, we need input from the phonetic research communities to identify suitable research targets.

Acknowledgements: ZM and GEH thank Jens Edlund for helpful discussions. Grant support for ZM, JB: Swedish Research Council no. 2017-02861; GEH: Swedish Foundation for Strategic Research no. RIT15-0107; CVB and OW: EPSRC Standard Research Grant EP/P011586/1; JG: Swedish Research Council no. 2013-4935.

5. REFERENCES

[1] Anumanchipalli, G. K., Cheng, Y.-C., Fernandez, J., Huang, X., Mao, Q., Black, A. W. 2010. KlaTTStat:

Knowledge-based parametric speech synthesis. *Proc. SSW* volume 7.

[2] Baayen, R. H., Milin, P. 2010. Analyzing reaction times. *Int. J. Psychol. Res.* 3(2), 12–28.

[3] Blomert, L., Mitterer, H. 2004. The fragile nature of the speech-perception deficit in dyslexia: Natural vs. synthetic speech. *Brain Lang.* 89(1), 21–26.

[4] Carlson, R., Granström, B., Hunnicutt, S. 1982. A multilanguage text-to-speech module. *Proc. ICASSP* 1604–1607.

[5] Cerňak, M., Beňuš, Š., Lazaridis, A. 2017. Speech vocoding for laboratory phonology. *Comput. Speech Lang.* 42, 100–121.

[6] Cooke, M., Mayo, C., Valentini-Botinhao, C. 2013. Intelligibility-enhancing speech modifications: The Hurricane Challenge. *Proc. Interspeech* 3552–3556.

[7] Duffy, S. A., Pisoni, D. B. 1992. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Lang. Speech* 35(4), 351–389.

[8] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vrecken, O. 1996. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proc. ICSLP* 1393–1396.

[9] Espic, F., Valentini-Botinhao, C., King, S. 2017. Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis. *Proc. Interspeech* 1383–1387.

[10] Griffin, D., Lim, J. 1984. Signal estimation from modified short-time fourier transform. *IEEE T. Acoust. Speech* 32(2), 236–243.

[11] Henter, G. E., Lorenzo-Trueba, J., Wang, X., Kondo, M., Yamagishi, J. 2018. Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody. *Proc. ICASSP* 4799–4803.

[12] Henter, G. E., Lorenzo-Trueba, J., Wang, X., Yamagishi, J. 2018. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. *arXiv preprint arXiv:1807.11470*.

[13] House, A. S., Williams, C., Hecker, M. H. L., Kryter, K. D. 1963. Psychoacoustic speech tests: A modified rhyme test. *J. Acoust. Soc. Am.* 35(11), 1899–1899.

[14] International Telecommunication Union, Radiocommunication Sector Oct. 2015. *Method for the subjective assessment of intermediate quality levels of coding systems*.

[15] Iverson, P. 2003. Evaluating the function of phonetic perceptual phenomena within speech recognition: An examination of the perception of /d/–/t/ by adult cochlear implant users. *J. Acoust. Soc. Am.* 113(2), 1056–1064.

[16] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., et al., 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Proc. NeurIPS* 4485–4495.

[17] King, S. 2014. Measuring a decade of progress in text-to-speech. *Loquens* 1(1), e006.

[18] King, S. 2015. What speech synthesis can do for you (and what you can do for speech synthesis). *Proc. ICPhS*.

[19] King, S., Karaiskos, V. 2013. The Blizzard Challenge 2013. *Proc. Blizzard Challenge Workshop*.

[20] Liberman, A. M., Mattingly, I. G. 1985. The motor theory of speech perception revised. *Cognition* 21(1), 1–36.

[21] Lisker, L., Abramson, A. S. 1970. The voicing dimension: Some experiments in comparative phonetics. *Proc. ICPhS* 563–567.

[22] Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., Ling, Z.-H. 2018. The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods. *Proc. Odyssey Workshop* 195–202.

[23] Luong, H.-T., Takaki, S., Henter, G. E., Yamagishi, J. 2017. Adapting and controlling DNN-based speech synthesis using input codes. *Proc. ICASSP* 4905–4909.

[24] Malisz, Z., Berthelsen, H., Beskow, J., Gustafson, J. 2017. Controlling prominence realisation in parametric DNN-based speech synthesis. *Proc. Interspeech* 1079–1083.

[25] Moulines, E., Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* 9(5-6), 453–467.

[26] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., et al., 2016. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[27] Ribeiro, M. S., Yamagishi, J., Clark, R. A. J. 2015. A perceptual investigation of wavelet-based decomposition of $f0$ for text-to-speech synthesis. *Proc. Interspeech* 1586–1590.

[28] Richmond, K., Clark, R. A. J., Fitt, S. 2009. Robust LTS rules with the Combilex speech technology lexicon. *Proc. Interspeech* 1295–1298.

[29] Richmond, K., Clark, R. A. J., Fitt, S. 2010. On generating Combilex pronunciations via morphological analysis. *Proc. Interspeech* 1974–1977.

[30] Rothauser, E. H., et al., 1969. IEEE recommended practice for speech quality measurements. *IEEE T. Acoust. Speech* 17(3), 225–246.

[31] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., et al., 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. *Proc. ICASSP* 4799–4783.

[32] Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., Granström, B. 1998. Web-based educational tools for speech technology. *Proc. ICSLP*.

[33] Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., et al., 2018. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. *Proc. ICML* 4693–4702.

[34] Tachibana, H., Uenoyama, K., Aihara, S. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *Proc. ICASSP* 4784–4788.

[35] Taylor, P., Black, A. W., Caley, R. 1998. The architecture of the Festival speech synthesis system. *Proc. SSW* volume 3 147–152.

[36] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., et al., 2017. Tacotron: Towards end-to-end speech synthesis. *Proc. Interspeech* 4006–4010.

[37] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., et al., 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *Proc. ICML* 5180–5189.

[38] Watts, O., Henter, G. E., Merritt, T., Wu, Z., King, S. 2016. From HMMs to DNNs: where do the improvements come from? *Proc. ICASSP* 5505–5509.

[39] Winters, S. J., Pisoni, D. B. 2004. Perception and comprehension of synthetic speech. *Research on Spoken Language Processing Progress Report* (26), 95–138.

[40] Wu, Z., Watts, O., King, S. 2016. Merlin: An open source neural network speech synthesis system. *Proc. SSW* volume 9 218–223.

[41] Xu, Y., Prom-On, S. 2014. Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Commun.* 57, 181–208.